ISOMORPHISM AND SYMMETRIES IN RANDOM PHYLOGENETIC TREES

MIKLÓS BÓNA AND PHILIPPE FLAJOLET

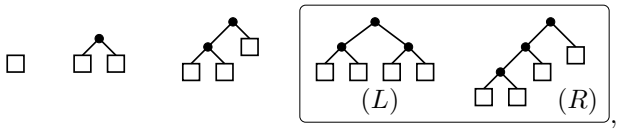
Abstract. The probability that two randomly selected phylogenetic trees of the same size are isomorphic is found to be asymptotic to a decreasing exponential modulated by a polynomial factor. The number of symmetrical nodes in a random phylogenetic tree of large size obeys a limiting Gaussian distribution, in the sense of both central and local limits. The probability that two random phylogenetic trees have the same number of symmetries asymptotically obeys an inverse square-root law. Precise estimates for these problems are obtained by methods of analytic combinatorics, involving bivariate generating functions, singularity analysis, and quasi-powers approximations.

1. Introduction

Every high school student of every civilized part of the world is cognizant of the *tree of species*, also known as the "tree of life", in relation to Darwin's theory of evolution (Figure 1). We observe n different species, and form a group with the closest pair (under some suitable proximity criterion), then repeat the process with the $n-2$ remaining species together with the newly formed group, and so on. In this way a *phylogenetic tree*, also known as "cladogram", is obtained: such a tree has the n species at its external nodes, also called "leaves"; it has $n-1$ internal binary nodes, and it is naturally rooted at the last node obtained by the process. Note that, by design, there is no specified order between the two children of a binary node.

Seen from combinatorics, the phylogenetic trees under consideration are thus trees in the usual sense of graph theory (i.e., acyclic connected graphs [4, §1.5]); in addition, a binary node is distinguished as the root, and each node has outdegree either 0 (leaf) or 2 (internal binary node). Finally, the leaves are labeled by distinct integers, which we may canonically take to be an integer interval $[1, n]$. In classical combinatorial terms, the set of phylogenetic trees thus corresponds to the set $\mathcal{B}$ of *rooted non-plane binary trees*, which are *labeled at their leaves*.

We let $\mathcal{B}_n$ be the subset of $\mathcal{B}$ corresponding to trees of size n (those with n leaves) and denote by $b_n := |\mathcal{B}_n|$ the corresponding cardinality. Considering the listing of all unlabeled trees of sizes $1, 2, 3, 4$

(1) 

,

the reader is invited to verify that $b_1 = 1$, $b_2 = 1$, $b_3 = 3$, and that $b_4 = 15$ is obtained by counting all possible labelings (3 and 12, respectively) of the two trees L, R shown on the right of (1).

A general formula for the numbers b_n is well known and straightforward to prove. Indeed, if we introduce the *exponential generating function*

$$B(z) := \sum_{n\geq 1} b_n \frac{z^n}{n!},$$

Date: January 6, 2009.

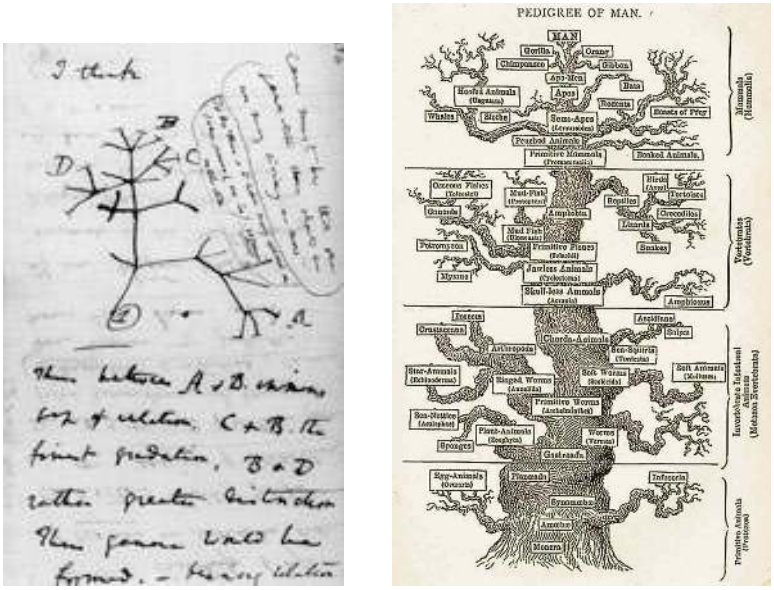


FIGURE 1. Left: the representation of a pylogenetic tree in Darwin's own handwriting. Right: an illustration of the Tree of Life by Haeckel in *The Evolution of Man*, published in 1879. (Source: Entry "Tree of life", *Wikipedia*.)

then the fact that each element of $\mathcal{B}_n$ is built up from its two subtrees implies that

$$B(z) = z + \frac{1}{2}B(z)^2. \tag{2}$$

See the books by Stanley [20, pp. 13–15] or Flajolet–Sedgewick [9, §2.5] for details and related results. So, $B(z)$ is the solution of the quadratic equation (2) that is a generating function. That is,

$$B(z) = 1 - \sqrt{1-2z}.$$

This leads to the following exact formula for the numbers b_n.

Proposition 1. *The number of phylogenetic trees on n labeled nodes is*

$$b_n = 1 \cdot 3 \cdot \cdots (2n-3) \equiv (2n-3)!!.$$

There is a natural way to associate an *unlabeled* rooted binary non-plane tree to each element $t \in \mathcal{B}_n$, by simply removing all the labels of t. We will say that two elements $t, t' \in \mathcal{B}_n$ are *isomorphic* if removing their labels will associate them to the same unlabeled tree. This leads to the following intriguing question.

> **Question.** *What is the probability p_n that two phylogenetic trees, selected uniformly at random in $\mathcal{B}_n$, are isomorphic?*

Note that, in our running example, the case of $n = 4$, we have $p_4 = \left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2 = \frac{17}{25}$. Indeed, if we selected two elements of $\mathcal{B}_4$ at random, there is a $(3/15)^2 = (1/5)^2$ chance that they will both belong to the isomorphism class of L, and $(12/15)^2 = (4/5)^2$ that they both belong to the isomorphism class of R, where L and R are the two trees of (1).

In this paper, we will use a multivariate generating function argument (Section 2) in conjunction with an analysis of singularities in the complex plane (Section 3) to answer the isomorphism question in Theorem 1. In Section 4, we will extend our analysis to distributional estimates of the number of symmetrical nodes in phylogenetic trees and in their unlabeled counterparts, known as Otter trees: see Theorems 2 and 3 for *central and local limit laws*, respectively. Such results in particular quantify the distribution of the log-size of the automorphism group of the random trees under consideration. In Section 5, we will work out an explicit estimate of the probability that two random trees have the same number of symmetries.

2. Isomorphism: a Generating Function Argument

2.1. **Unlabeled Trees.** Let $\mathcal{U}_n$ be the set of all *unlabeled* rooted binary non-plane trees with n leaves, and let $u_n = |\mathcal{U}_n|$ be the corresponding count, with *ordinary generating function*

$$U(z) := \sum_{n\geq 1} u_n z^n.$$

Such trees are often called *Otter trees*, since Otter was the first to study their enumeration [17]. We can build a generic element of $\mathcal{U}_n$ by taking a tree $t' \in \mathcal{U}_k$ and a tree $t'' \in \mathcal{U}_{n-k}$, and joining their roots to a new root. As the order of t and t' is not significant, we get each tree $t \in \mathcal{U}_n$ *twice* this way, except that, if the two subtrees of t are identical, we get t only once. This leads to the functional equation [9, 12, 17, 18]:

$$U(z) = z + \frac{1}{2}\left(U(z)^2 + U(z^2)\right). \tag{3}$$

The numbers u_n are listed as sequence A001190 (the *"Wedderburn–Etherington numbers"*) in the On-line Encyclopedia of Integer Sequences by Neil Sloane [19] and are the answers to various combinatorial enumeration problems. The first few values of the sequence $\{u_n\}_{n\geq 1}$ are 1, 1, 1, 2, 3, 6, 11, 23, 46, 98.

2.2. **A multivariate generating function.** Let $t_1 \in \mathcal{B}_n$, and let $t_2 \in \mathcal{B}_n$. By Proposition 1, there are $(2n-3)!!^2$ possibilities for the ordered pair (t_1, t_2), where t_1 and t_2 do not have to be distinct. Our goal is to count such ordered pairs in which t_1 and t_2 are isomorphic. This number, divided by $(2n-3)!!^2$ will then provide the probability p_n that two randomly selected elements of $\mathcal{B}_n$ are isomorphic.

Let $t \in \mathcal{U}_n$. Then the number of different labelings of the leaves of t is

$$w(t) = \frac{n!}{2^{\mathrm{sym}(t)}}, \tag{4}$$

where $\mathrm{sym}(t)$ is the number of non-leaf nodes v of t such that the two subtrees stemming from v are identical. For example, if $n = 4$, and t is the tree L of (1), then we have $w(t) = 3$, and indeed, t has $n!/2^3 = 24/8 = 3$ labelings. If t is the tree R of (1), then we have $w(t) = 1$, and t has $24/2 = 12$ labelings.

Isomorphism classes within $\mathcal{B}_n$ correspond to elements of $\mathcal{U}_n$. Set

$$W_n = \sum_{t\in\mathcal{U}_n} \frac{1}{2^{\mathrm{sym}(t)}}. \tag{5}$$

As we have mentioned above, $n!/2^{\mathrm{sym}(t)}$ is the number of labeled trees in the isomorphism class corresponding to t. Summing this number over all isomorphism

classes, we obtain the total number of trees in $\mathcal{B}_n$. That is,

$$n!W_n = 1 \cdot 3 \cdot \cdots \cdot (2n-3)!!.$$

For instance, $W_4 = \frac{1}{8} + \frac{1}{2} = \frac{5}{8}$, and $4! \cdot \frac{5}{8} = 15 = 5!!$.

Let

$$F(z,u) = \sum_{t \in \mathcal{U}} u^{\mathrm{sym}(t)} z^{|t|} \tag{6}$$

be the bivariate generating function of Otter trees, with z marking the number of leaves, and u marking non-leaf nodes with two identical subtrees. In particular, $F(z,u) = z + uz^2 + uz^3 + (u^3+u)z^4 +$ higher degree terms. The crucial observation about $F(z,u)$ is the following.

Lemma 1. *The bivariate generating function $F(z,u)$ that enumerates Otter trees with respect to the number of symmetrical nodes satisfies the functional equation*

$$F(z,u) = z + \frac{1}{2}F(z,u)^2 + \left(u - \frac{1}{2}\right) F(z^2,u^2). \tag{7}$$

Proof. If a tree consists of more than one node, then it is built up from its two subtrees. As the order of the two subtrees is not significant, we will get each tree *twice* this way, except the trees whose two subtrees are identical. If t_1 and t_2 are the two subtrees of t whose roots are the two children of the root of t, then

$$\mathrm{sym}(t) = \begin{cases} \mathrm{sym}(t_1) + \mathrm{sym}(t_2), & \text{if } t_1 \text{ and } t_2 \text{ are not identical} \\ \mathrm{sym}(t_1) + \mathrm{sym}(t_2) + 1, & \text{if } t_1 \text{ and } t_2 \text{ are identical.} \end{cases}$$

The first term of the right-hand side of (7) represents the tree on one node, the second term represents all other trees as explained in the preceding paragraph, and the third term is the correction term for trees in which the two subtrees of the root are identical. □

Note that various specializations of $F(z,u)$ have a known combinatorial meaning. Indeed,

(*i*) If $u = 1$, then $F(z,1) = \sum_{t \in U} z^{|t|}$ is simply the ordinary generating function $U(z)$ of Otter trees with respect to their number of leaves. We have discussed this generating function in Subsection 2.1, and mentioned that its coefficients u_n are the Wedderburn–Etherington numbers, which form sequence A001190 in [19].

(*ii*) If $u = 2$, then $F(z,2) = \sum_{t \in U} z^{|t|} 2^{\mathrm{sym}(t)}$ is the ordinary generating function of the total number of automorphisms in all Otter trees. The coefficients constitute sequence A003609 in [19]. Interested readers may consult McKeon's studies [14, 15] for details. The first few elements of the sequence are 1, 2, 2, 10, 14, 42, 90, 354.

(*iii*) If $u = 1/2$, then

$$F\left(z, \frac{1}{2}\right) = \sum_{t \in \mathcal{U}} z^{|t|} 2^{-\mathrm{sym}(t)} = \sum_n W_n z^n = \sum_n (2n-3)!! \frac{z^n}{n!},$$

is the exponential generating function $B(z)$ of labeled trees in disguise. We have discussed this generating function in the Introduction. The numbers $(2n-3)!!$ form sequence A001147 in [19].

It is more surprising that the substitution $u = 1/4$ will give us the answer we are seeking. Let $[z^n]g(z)$ denote the coefficient of z^n in the power series $g(z)$.

Lemma 2. *For all positive integers $n \geq 2$, the probability p_n that two phylogenetic trees of size n are isomorphic satisfies*

$$p_n = \left(\frac{n!}{(2n-3)!!}\right)^2 \cdot [z^n] F\left(z, \frac{1}{4}\right).$$

Proof. Consider the sample space whose elements are the elements of $\mathcal{U}_n$, and in which the probability of $t \in \mathcal{U}_n$ is

$$\kappa(t) := \frac{n!}{2^{\mathrm{sym}(t)}} \cdot \frac{1}{(2n-3)!!} = \frac{w(t)}{(2n-3)!!}. \tag{8}$$

(For probabilists, κ is the image on $\mathcal{U}_n$ of the uniform distribution of $\mathcal{B}_n$.) For instance, if $n = 4$, then this space has two elements, (the two trees L, R of (1)), one has probability $1/5$, and the other has probability $4/5$. If we select two elements of this space at random, the probability that they coincide is

$$p_n = \sum_{t \in \mathcal{U}_n} \kappa(t)^2 = \frac{1}{(2n-3)!!^2} \sum_{t \in \mathcal{U}_n} w(t)^2 = \frac{n!^2}{(2n-3)!!^2} \sum_{t \in \mathcal{U}_n} \left(\frac{1}{4}\right)^{\mathrm{sym}(t)}.$$

Our claim now follows since $\sum_{t \in \mathcal{U}_n} \left(\frac{1}{4}\right)^{\mathrm{sym}(t)}$ is indeed the coefficient of z^n in $F(z, 1/4)$, in accordance with the definition (6). □

3. Isomorphism: Singularity Analysis

By Lemma 2, our goal is now to find the coefficient of z^n in the one-variable generating function

$$f(z) := F(z, 1/4).$$

Lemma 1 shows that the formal power series $F(z, u)$ is the solution of the quadratic equation (7) that satisfies $F(0, 0) = 0$. That is,

$$F(z, u) = 1 - \sqrt{1 - 2z - (2u-1)F(z^2, u^2)}. \tag{9}$$

Iterated applications of (9), starting with $u = 1/4$, show that

$$\begin{aligned} f(z) \;\equiv\; F(z, 1/4) \;&=\; 1 - \sqrt{1 - 2z + \frac{1}{2} F\left(z^2, \frac{1}{16}\right)} \\ &= 1 - \sqrt{\frac{3}{2} - 2z - \frac{1}{2}\sqrt{1 - 2z^2 + \frac{7}{8} F\left(z^4, \frac{1}{256}\right)}} \;=\; \cdots. \end{aligned}$$

In the limit, there results that $f(z)$ admits a "continued square-root" expansion

$$f(z) = 1 - \sqrt{\frac{3}{2} - 2z - \frac{1}{2}\sqrt{\frac{15}{8} - 2z^2 - \frac{7}{8}\sqrt{\frac{255}{128} - 2z^4 - \frac{127}{128}\sqrt{\cdots}}}},$$

out of which initial elements of the sequence $(p_n)_{n \geq 1}$ are easily determined:

$$1,\ 1,\ 1,\ \frac{17}{25},\ \frac{3}{7},\ \frac{5}{21},\ \frac{13}{99},\ \frac{1385}{20449},\ \frac{17861}{511225},\ \frac{101965}{5909761},\ \cdots.$$

In order to compute the growth rate of the coefficients of $f(z)$, we will analyze the dominant singularity (or singularities) of this power series. The interested reader is invited to consult the book *Analytic Combinatorics* by Flajolet and Sedgewick [9] for more information on the notions and techniques that we are going to use. Part

of the difficulty of the problem is that the functional relation (9) has the character of an inclusion–exclusion formula: $F(z,u)$ does *not* depend positively on $F(z^2,u^2)$, as soon as $u \leq 1/2$, which requires suitably crafted arguments, in contrast to the (simpler) asymptotic analysis of $u_n = [z^n]F(z,1)$.

Briefly, we are interested in the *location*, *type*, and *number* of the dominant singularities of $f(z)$, that is, singularities that have smallest absolute value (modulus).

3.1. **Location.** First, it is essential for our analytic arguments to establish that $f(z)$ has a radius of convergence strictly less than 1. Our starting point parallels Lemmas 1–2 of McKeon [15], but we need a specific argument for the upper bound.

Lemma 3. *Let ρ be the largest real number such that $f(z)$ is analytic in the interior of a disc centered at the origin that has radius ρ. The following inequalities hold:*

$$0.4 < \rho < 0.625.$$

Proof. (*i*) *Lower bound.* Note that $f(z)$ is convergent in some disc of radius *at least* 0.4, since the coefficients of $f(z) = F(z,1/4)$ are at most as large as the coefficients of $F(z,1)$, the generating function $U(z)$ of Otter trees, and the latter is known to be convergent in a disc of radius $0.40269\cdots$: see Otter's original paper [17] and Finch's book [6, §5.6] for more details on the asymptotics of $F(z,1) = U(z)$.

(*ii*) *Upper bound.* For fixed n, let $a_1, a_2, \cdots, a_{u_n}$ be the numbers of our labeled trees whose underlying unlabeled tree is the first, second, ..., last Otter tree of size n. Then the relation

$$p_n \equiv \frac{a_1^2 + a_2^2 + \cdots + a_{u_n}^2}{(a_1 + a_2 + \cdots + a_{u_n})^2} \;>\; \frac{1}{u_n}, \tag{10}$$

results from the Cauchy-Schwarz inequality. (In words: the probability of coincidence of two elements from a finite probability space is smallest when the distribution is the uniform one.)

As we mentioned, it is proved in [17] that the generating function $\sum_n u_n x^n$ converges in a disc of radius at least 0.4. Therefore, the series $\sum_n \frac{1}{u_n} x^n$ converges in a disc of radius at most $1/0.4 = 2.5$, and by (10), this implies that $\sum_n p_n x^n$ converges in a disc of radius less than 2.5. Now Lemma 2 shows that $F(z,1/4)$ is convergent in a disc of radius less than $2.5/4 = 0.625$, since the coefficients of $F(z,1/4)$ are, up to polynomial factors, 4^n times larger than the coefficients of $\sum_n p_n x^n$. It follows that $\rho < 0.625$. □

A well-known theorem of Pringsheim states that if a function $g(z)$ is representable around the origin by a series expansion that has non-negative coefficients and radius of convergence R, then the real number R is actually a singularity of $g(z)$. Applying this theorem to $f(z)$, we see that the positive real number ρ must be a singularity of $f(z)$.

3.2. **Type.** Recall that a function $g(z)$ analytic in a domain Ω is said to have a *square-root singularity* at a boundary point α if, for some function H analytic at 0, the representation $g(z) = H(\sqrt{z-\alpha})$ holds in the intersection of Ω and a neighborhood of α. (In particular, if $g(z) = \sqrt{\gamma(z)}$ with γ analytic at α, then $g(z)$ has a square-root singularity at α whenever $\gamma(\alpha) = 0$ and $\gamma'(\alpha) \neq 0$.)

Lemma 4. *All dominant singularities (of modulus ρ) of $f(z)$ are isolated and are of the square-root type.*

Proof. In order to see this, note that $\rho < 1$ (proved in Lemma 3) implies that $\rho < \sqrt{\rho}$. Therefore, the power series $F(z^2, 1/4)$ (that has radius of convergence $\sqrt{\rho}$) is analytic in the interior of the disc of radius ρ, and so is the power series $F(z^2, 1/16)$ since its coefficients are smaller than the corresponding coefficients of $F(z^2, 1/4)$. Consequently, Equation (9) implies that the dominant singularities of

$$f(z) = F\left(z, \frac{1}{4}\right) = 1 - \sqrt{1 - 2z + \frac{1}{2}F\left(z^2, \frac{1}{16}\right)} \tag{11}$$

are of the square-root type: they are to be found amongst the roots of the expression under the square-root sign in (9), that is, amongst the zeros of $1-2z+\frac{1}{2}F(z^2, 1/16)$ that have modulus ρ. As $1 - 2z + \frac{1}{2}F(z^2, 1/16)$ is analytic in the disc centered at the origin with radius at least $\sqrt{\rho} > \rho$, it has isolated roots. Hence $f(z)$ has only a *finite* number of singularities on the circle $|z| = \rho$, and each is of square-root type. □

The argument of the proof (see (11)) also shows that ρ is determined as the smallest positive root of the equation

$$1 - 2\rho + \frac{1}{2}F\left(\rho^2, \frac{1}{16}\right) = 0. \tag{12}$$

3.3. **Number.** In order to complete our characterization of the dominant singular structure of $f(z)$, we need the following statement.

Lemma 5. *The point ρ is the* only *singularity of smallest modulus of $f(z)$.*

Proof. The argument is somewhat indirect and it proceeds in two stages.

First we show that, as a power series, $f(z)$ converges for each z with $|z| = \rho$. To this purpose, we need to recall briefly some principles of singularity analysis, as expounded in [9, Ch. VI]. Let $g(z)$ be a function analytic in $|z| < R$ with finitely many singularities at the set $\{\alpha_j\}$ on the circle $|z| = R$; assume in addition that $g(z)$ has a square-root singularity at each α_j in the sense of Subsection 3.2. Then, one has $[z^n]g(z) = O\left(R^{-n}n^{3/2}\right)$. (This corresponds to the O–transfer theorem of [9, Th. VI.3, p. 390], with amendments for the case of multiples singularities to be found in [9, §VI.5]; see also (14) below.) It follows from this general estimate and Lemma 4 that

$$[z^n]f(z) = O(\rho^{-n}n^{3/2}).$$

Therefore, the series expansion of $f(z)$ converges absolutely as long as $|z| \leq \rho$, and, in particular, it converges for all z with modulus ρ.

Now, we are in a position to prove that $f(z)$ has no singularity other than ρ on the circle $|z| = \rho$. Let us assume the contrary; that is, there is a real number $z_0 \neq \rho$ such that $|z_0| = \rho$ and z_0 is a singularity of $f(z) \equiv F(z, 1/4)$. Then, it follows from (9) that $f(z_0) \equiv F(z_0, 1/4) = 1$, since the expression under the square-root sign in (9) is equal to 0, corresponding to a singularity of square-root type. On the other hand, one has a priori $|f(z_0)| \leq f(\rho)$, as a consequence of the triangle inequality and the fact, proved above, that $f(z)$ converges on $|z| = \rho$. Now it follows from the *strong triangle inequality* that the equality $f(z_0) = f(\rho)$ is only possible if all the terms $f_n z_0^n$ that compose the (convergent) series expansion of $f(z_0)$ are positive real. (Here $f_m = [z^m]f(z)$.) However, since, in particular, $f_1 = 1$ is nonzero, this implies that $z_0 = \rho$, and a contradiction has been reached. (This part of the argument is also closely related to the Daffodil Lemma of [9, p. 266].) □

3.4. **The asymptotics of p_n.** As a result of Lemmas 3–5, the function $f(z)$ has only one dominant singularity, and that singularity ρ is of the square-root type. One then has, for a family of constants h_k, the local singular expansion:

$$f(z) = 1 + \sum_{k=0}^{\infty} h_k (1 - z/\rho)^{k+1/2}, \tag{13}$$

which is valid for z near ρ. The conditions of the singularity analysis process as summarized in [9, §VI.4] are then satisfied. Consequently, each singular element of (13) relative to $f(z)$ can be translated into a matching asymptotic term relative to $[z^n]f(z)$, according to the rule

$$\sigma(z) = (1 - z/\rho)^{\theta} \quad \longrightarrow \quad [z^n]\sigma(z) = \rho^{-n} \binom{n-\theta-1}{n} \sim \rho^{-n} \frac{n^{-\theta-1}}{\Gamma(-\theta)}. \tag{14}$$

In particular, we have $[z^n]f(z) \sim C \cdot \rho^{-n} n^{-3/2}$, for some C.

Hence Lemma 2, combined with Lemmas 4–5 and the routine asymptotics of $n!/(2n-3)!!$ by Stirling's formula, leads to the following theorem.

Theorem 1. *The probability that two phylogenetic trees of size n are isomorphic admits a complete asymptotic expansion*

$$p_n \sim a \cdot b^{-n} \cdot n^{3/2} \left(1 + \sum_k \frac{c_k}{n^k} \right), \tag{15}$$

where a, $b = 4\rho$, and the c_k are computable constants, with values $a = 3.17508\cdots$, $b = 2.35967\cdots$, and c_1 approximately equal to -0.626.

The function $F(z,u)$ can be determined numerically to great accuracy (by means of the recursion corresponding to the functional equation (9)). So, the value

$$\rho = 0.58991\,82714\,85535\cdots,$$

is obtained as the smallest positive root of (12); the constant a then similarly results from an evaluation of $F'\left(\rho^2, \frac{1}{16}\right)$; the constant c_1, which could in principle be computed in the same manner, was, in our experiments, simply estimated from the values of p_n for small n. The formula (15), truncated after its c_1/n term, then appears to approximate p_n with a relative accuracy better than 10^{-2} for $n \geq 5$, 10^{-4} for $n \geq 38$, and 10^{-5} for $n \geq 47$.

4. Symmetrical Nodes and Automorphisms

In the course of our investigations on analytic properties of the bivariate generating function $F(z,u)$, we came up with a few additional estimates, which improve on those of McKeon [15]. In essence, what is at stake is a perturbative analysis of $F(z,u)$ and its associated singular expansions, for various values of u, in a way that refines the developments of the previous section. We offer here a succinct account: details can be easily supplemented by referring to Chapter IX of the book *Analytic Combinatorics* [9].

Theorem 2. (*i*) *Let X_n be the random variable representing the number of symmetrical nodes in a random Otter tree of $\mathcal{U}_n$. Then, X_n satisfies a limit law of Gaussian type,*

$$\forall x \in \mathbb{R} \ : \qquad \lim_{n\to\infty} \mathbb{P}\left(X_n \leq \mu n + \sigma x \sqrt{n}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-w^2/2}\, dw,$$

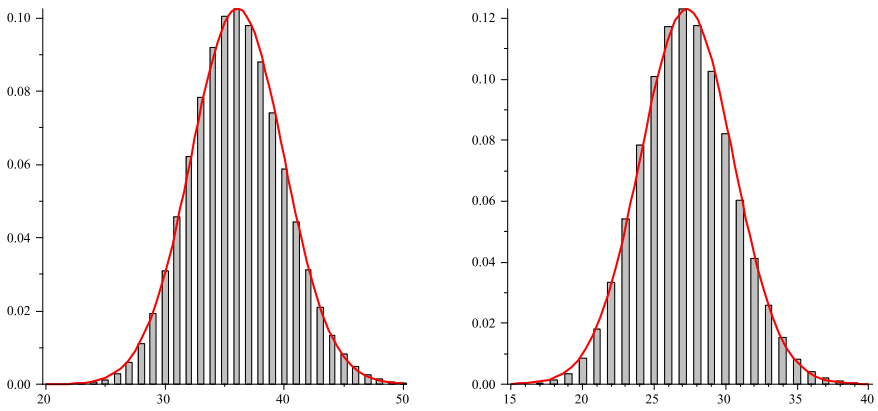


FIGURE 2. Histograms of the distribution of the number of symmetrical nodes in trees of size 100, compared to a matching Gaussian. Left: Otter trees of $\mathcal{U}_{100}$. Right: phylogenetic trees of $\mathcal{B}_{100}$.

for some positive constants μ *and* σ. *Numerically,* $\mu = 0.35869\cdots$.

(*ii*) *Let* Y_n *be the random variable representing the number of symmetrical nodes in a random phylogenetic tree of* $\mathcal{B}_n$. *Then,* Y_n *satisfies a limit law of Gaussian type,*

$$\forall x \in \mathbb{R} \ : \qquad \lim_{n\to\infty} \mathbb{P}\left(Y_n \le \widehat{\mu} n + \widehat{\sigma} x \sqrt{n}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-w^2/2}\, dw,$$

for some positive constants $\widehat{\mu}$ *and* $\widehat{\sigma}$. *Numerically,* $\widehat{\mu} = 0.27104\cdots$.

Proof (Sketch). (*i*) *The case of Otter trees* $(X_n, \mathcal{U}_n)$. In accordance, with general principles [9, Ch. IX], we need to estimate the generating polynomial

$$\varphi_n(u) := [z^n] F(z, u), \tag{16}$$

when u is close to 1, with $F(z, u)$ as specified by (6) and (7). For u in a small enough complex neighborhood Ω of 1, the radius of convergence of $F(z^2, u^2)$ is larger than some $\rho_2 > \rho_1$, where $\rho_1 \approx 0.40269$ is the radius of convergence associated with Otter trees. Then, by an argument similar to the ones used earlier, there exists a solution $\rho(u)$ to the analytic equation

$$1 - 2\rho(u) + (u-1) F(\rho(u)^2, u^2) = 0 \tag{17}$$

(compare with (12)), such that $\rho(1) = \rho_1$ is the dominant singularity of the generating function $F(z, 1)$ of Otter trees. By the analytic version of the implicit function theorem (equivalently, by the Weierstrass Preparation Theorem), this function $\rho(u)$ depends analytically on u, for u near 1.

In addition, by (9), the function $F(z, u)$ has a singularity of the square-root type at $\rho(u)$. Also, for $u \in \Omega$ and Ω taken small enough, the triangle inequality combined with the previously established properties of $F(z, 1)$ may be used to verify that there are no other singularities of $z \mapsto F(z, u)$ on $|z| = |\rho(u)|$. There results, from singularity analysis *and* the uniformity of the process [9, p. 668], the asymptotic estimate

$$\varphi_n(u) = c(u) \rho(u)^{-n} n^{-3/2} \left(1 + o(1)\right), \qquad n \to +\infty, \tag{18}$$

uniformly with respect to $u \in \Omega$, for some $c(u)$ that is analytic at $u = 1$. Then, the probability generating function of X_n, which equals $\varphi_n(u)/\varphi_n(1)$ satisfies what is known as a *"quasi-powers approximation*. That is, it resembles (analytically) the probability generating function of a sum of independent random variables,

$$\frac{\varphi_n(u)}{\varphi_n(1)} = \frac{c(u)}{c(1)} \left(\frac{\rho(1)}{\rho(u)} \right)^n [1 + \varepsilon_n(u)], \tag{19}$$

where $\sup_{u\in\Omega} |\varepsilon_n(u)|$ tends to 0 as $n \to \infty$. The Quasi-powers Theorem (see [9, §IX.5] and [13]) precisely applies to such approximations by quasi-powers and implies that the distribution of X_n is asymptotically normal.

(*ii*) *The case of phylogenetic trees* $(Y_n, \mathcal{B}_n)$. The starting point is a simple combinatorial property of $\varphi_n(u)$, as defined in (16):

$$\varphi_n(u/2) = \frac{1}{n!} \sum_{t\in\mathcal{U}_n} \frac{n!}{2^{\mathrm{sym}(t)}} u^{\mathrm{sym}(t)} = \frac{1}{n!} \sum_{t\in\mathcal{B}_n} u^{\mathrm{sym}(t)}. \tag{20}$$

(The first form results from the definition (6) of $F(z, u)$; the second form relies on the expression (4) of the number of different labellings of an Otter tree that give rise to a phylogenetic tree.) Thus, φ_n taken with an argument near 1/2 serves, up to normalization, as the probability generating function of the number of symmetrical nodes in phylogenetic trees of $\mathcal{B}_n$.

From this point on, the analysis of symmetries in phylogenetic trees is entirely similar to that of Otter trees. For u in a small complex neighborhood $\widehat{\Omega}$ of 1/2, the generating function $z \mapsto F(z, u)$ has a dominant singularity $\rho(u)$ that is an analytic solution of (17) and is such that $\rho(1/2) = 1/2$, the radius of convergence of $B(z) \equiv F(z, 1/2)$. As a consequence, estimates that parallel those of (18) and (19) are seen to hold, but with $u \in \widehat{\Omega}$ now near 1/2. In particular,

$$\frac{\varphi_n(u)}{\varphi_n(1/2)} = \frac{\widehat{c}(u)}{\widehat{c}(1/2)} \left(\frac{\widehat{\rho}(1/2)}{\widehat{\rho}(u)} \right)^n [1 + \widehat{\varepsilon}_n(u)], \tag{21}$$

where $\widehat{\varepsilon}_n(u) \to 0$ uniformly. By the Quasi-powers Theorem (set $u := v/2$, with v near 1), the distribution of Y_n is asymptotically normal. □

Figure 2 shows that the fit with a Gaussian is quite good, even for comparatively low sizes ($n = 100$). Phrased differently, the statement of Theorem 2 means that the *logarithm of the order* $2^{\mathrm{sym}(t)}$ *of the automorphism group of a random tree* t *(either in* $\mathcal{U}_n$ *or in* $\mathcal{B}_n$*) is normally distributed*[1]. In the case of $\mathcal{U}_n$, the *expectation* of the cardinality of this group has been determined by McKeon [15] to grow roughly as 1.33609^n. In the case of phylogenetic trees $(\mathcal{B}_n)$, we find an *expected* growth of the rough form 1.24162^n, where the exponential rate $1.24162\cdots$ is exactly $1/(2\rho_1)$, with ρ_1, still, the radius of convergence of $U(z) \equiv F(z, 1)$. (These values are consistent with the fact that trees with a higher number of symmetries admit a smaller number of labellings, hence are less likely to appear as "shapes", under the phylogenetic model $\mathcal{B}_n$.)

As a matter of fact, the histograms of Figure 2 suggest that a convergence stronger than a plain convergence in law (corresponding to convergence of the distribution function) holds.

[1]The situation is loosely evocative of the fact (Erdős–Turán Theorem) that the logarithm of the order of a random permutation of size n is normally distributed; see, e.g., [5, 11, 16].

Definition 1. *Let (ξ_n) be a family of random variables with expectation $\mu_n = \mathbb{E}(\xi_n)$ and variance $\sigma_n^2 = \mathbb{V}(\xi_n)$. It is said to satisfy a local limit law with density $g(x)$ if one has*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\sigma_n \, \mathbb{P}(\xi_n = \lfloor \mu_n + x\sigma_n \rfloor) - g(x)| = 0. \tag{22}$$

In other terms, we expect the probability of ξ_n being at x standard deviations away from its mean to be well approximated by $g(x)/\sigma_n$. This concept is discussed in the case of sums of random variables by Gnedeneko and Kolmogorov in [10, Ch. 9] and, in a broader combinatorial context, by Bender [1] and Flajolet–Sedgewick [9, §IX.9].

Theorem 3. *The number of symmetrical nodes in either an unlabeled tree (X_n on $\mathcal{U}_n$) or a phylogenetic tree (Y_n on $\mathcal{B}_n$) satisfies a* local limit law *of the Gaussian type. That is, in the sense of Definition 1, a local limit law holds, with density*

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Proof. *(i) The unlabeled case* $(X_n, \mathcal{U}_n)$. The proof essentially boils down to establishing that

$$f_n(u) = [z^n] F(z, u)$$

is small compared to $[z^n]F(z,1)$, as soon as u satisfies $|u| = 1$ and stays away from 1; then, Theorem IX.14, p. 696, from [FlSe08] does the rest. The arguments are variations of the ones previously used.

Since a tree of size n has less than n symmetrical nodes, we have $|f_n(u)| \le |u|^n f_n(1)$ for any $|u| \ge 1$. There results that the convergence of the series expansion of $F(z,u)$ is dominated by that of $F(|zu|, 1)$, whenever $|u| \ge 1$. Apply the fact explained in the previous sentence, with z^2 and u^2 instead of z and u, to get that the coefficients of $F(z^2, u^2)$ are less than the coefficients of $F(|z^2u^2|, 1)$, where the latter series is convergent if $|z^2u^2| < 0.625$, or in other words, $|zu| < 0.75$, say. Now choose η so that $(1+\eta)(\rho_1 + \eta) < 0.75$, where ρ_1 is the radius of convergence of Otter trees ($\rho_1 \equiv \rho(1) \approx 0.40269$). Then $F(z^2, u^2)$ is bivariate analytic whenever $|z| < (\rho_1 + \eta)$ and $|u| < 1 + \eta$. In accordance with previously developed arguments, this implies that, for any fixed u satisfying $|u| \le 1 + \eta$, the function $z \mapsto F(z,u)$ has only finitely many singularities, each of the square-root type, in $|z| \le \rho_1 + \eta$.

For u in a small complex neighborhood of 1, we already know that $z \mapsto F(z,u)$ has only *one* dominant singularity at some $\rho(u)$, which is a root of

$$1 - 2\rho(u) + (2u-1) F(\rho(u)^2, u^2) = 0.$$

(This property lies at the basis of the central limit law of the previous theorem.)

Consider now a u such that $|u| = 1$, but $u \notin \Omega$. We argue that $z \mapsto F(z,u)$ is analytic at all points z such that $|z| = \rho_1$. Indeed for such values of u and z, we have, by the *strong triangle inequality*,

$$|F(z,u)| < F(\rho_1, 1), \tag{23}$$

the reason being that, in the expansion $F(z,u) = z + uz^2 + uz^3 + \cdots$, the values of the monomials $u^k z^n$ cannot be all collinear, unless $u = 1$. The inequality (23) combined with the fact that $F(\rho_1, 1) = 1$ implies that $z \mapsto F(z,u)$ cannot be singular (since, as we know, the only possibility for a singularity would be that it is of the square-root type *and* $F(z,u) = 1$).

Thus, for $|u| = 1$ and $u \notin \Omega$, the function $z \mapsto F(z, u)$ is analytic at all points of $|z| = \rho_1$. Hence, it is analytic in $|z| \leq \rho_1 + \delta$, for some $\delta > 0$. By usual exponential bounds, there results that, for some $K > 0$, one has

$$|f_n(u)| < K\,(\rho_1 + \delta/2)^{-n}\,, \qquad |u| = 1, \quad u \notin \Omega. \tag{24}$$

As expressed by Theorem IX.14[2] of [9], the existence of a quasi-powers approximation (when u is near 1), as in (18) and (19), and of the exponentially small bound (when $u \notin \Omega$ is away from 1), as provided by (24), suffices to ensure the existence of a local limit law.

(*ii*) *The labeled case* $(Y_n, \mathcal{B}_n)$. In accordance with (20), the function $F(z, u/2)$ is the bivariate exponential generating function of phylogenetic trees, with z marking size and u marking the number of symmetrical nodes. Consider once more $|u| = 1$ and distinguish the two cases $u \in \widehat{\Omega}$ (for which the proof of Theorem 2 provides a quasi-powers approximation) and $u \notin \widehat{\Omega}$. In the latter case, arguments that entirely parallel those applied to unlabeled trees give us that $z \mapsto F(z, u/2)$ has no singularity on $|z| = 1/2$. This implies, for $u \notin \widehat{\Omega}$, the exponential smallness of $\widehat{\varphi}_n(u/2)$, as defined in (20), resulting in an estimate that parallels (24). Theorem IX.14 of [9] again enables us to conclude as to the existence of a local limit law. □

5. Coincidence of the Number of Symmetries

From a statistician's point of view, it may be of interest to determine the probability for two trees to be "*similar*" (rather than plainly isomorphic), given some structural similarity distance between non-plane trees—see, for instance, the work of Ycart and Van Cutsem [21] for a study conducted under probabilistic assumptions that differ from ours. Combinatorial generating functions can still be useful in this broad range of problems, as we now show by considering the following question: *determine the probability that two randomly chosen trees τ, τ' of the same size have the same number of symmetrical nodes.* This probability *a priori* lies in the interval $[\frac{1}{n}, 1]$; we shall see, in Theorem 4, that its asymptotic value is "in-between".

The problem under consideration belongs to an orbit of questions occasionally touched upon in the literature. For instance, Wilf [22] showed that the probability that two permutations of size n have the same number of cycles is asymptotic to $(2\sqrt{\pi \log n})^{-1}$; Bóna and Knopfmacher [2] examine combinatorially and asymptotically the probability that various types of integer compositions have the same number of parts, and several other coincidence probabilities are studied in [7]. The following basic lemma trivializes the asymptotic side of several such questions.

Lemma 6. *Let $\mathcal{C}$ be a combinatorial class equipped with an integer-valued parameter χ. Assume that the random variable corresponding to χ restricted to $\mathcal{C}_n$ (under the uniform distribution over $\mathcal{C}_n$) satisfies a local limit law with density $g(x)$, in the sense of Definition 1. Let the variance of χ on $\mathcal{C}_n$ be σ_n^2 and assume that $g(x)$ is continuously differentiable. Then, the probability that two objects $c, c' \in \mathcal{C}_n$*

[2] The reasoning corresponding to that theorem is simple: start from

$$[u^k] f_n(u) = \frac{1}{2i\pi} \int_{|u|=1} f_n(u)\, \frac{du}{u^{k+1}}.$$

Use (24) to neglect the contribution corresponding to $u \notin \Omega$; appeal to the saddle point method applied to the quasi-powers approximation to estimate the central part $u \in \Omega$, and conclude.

admit the same value of χ satisfies the asymptotic estimate

$$\mathbb{P}\left[\chi(c)=\chi(c'),\quad c,c'\in\mathcal{C}_n\right]\sim\frac{K}{\sigma_n},\qquad \textit{where}\quad K:=\int_{-\infty}^{\infty}g(x)^2\,dx. \tag{25}$$

Note that, for $g(x)$ the standard Gaussian density, one has $K=1/(2\sqrt{\pi})$.

Proof (sketch). Let ϖ_n be the probability of coincidence; that is, the left hand-side of (25). Observe that, by hypothesis, we must have $\sigma_n\to\infty$. The baseline is that

$$\begin{aligned}\varpi_n &= \sum_k \mathbb{P}_{\mathcal{C}_n}[\chi(c)=k]^2\\ &\sim \frac{1}{\sigma_n^2}\sum_{x\in\mathcal{E}_n} g(x)^2, \qquad \text{with}\quad \mathcal{E}_n:=\frac{1}{\sigma_n}\left(\mathcal{Z}_{\geq 0}-\{\mu_n\}\right),\quad \mu_n:=\mathbb{E}_{\mathcal{C}_n}[\chi]\\ &\sim \frac{1}{\sigma_n}\int_{-\infty}^{\infty} g(x)^2\,dx.\end{aligned}$$

To justify this chain rigorously, first restrict attention to values of x in a finite interval $[-A,+B]$, so that the tails $(\int_{<A}+\int_{>B})g$ are less than some small ϵ. Then, with $x\in[-A,+B]$, make use of the approximation (22) provided by the assumption of a local limit law. Next, approximate the sum of $g(x)^2$ taken at regularly spaced sampling points (a Riemann sum) by the corresponding integral. Finally, complete back the tails. □

Given the local limit law expressed by Theorem 3, an immediate consequence of Lemma 6 is the following.

Theorem 4. *For Otter trees $(\mathcal{U}_n)$ and phylogenetic trees $(\mathcal{B}_n)$, the asymptotic probabilities that two trees of size n have the same number of symmetries admit the forms*

$$\mathcal{U}_n\ :\quad \frac{1}{2\sigma\sqrt{\pi n}},\qquad \mathcal{B}_n\ :\quad \frac{1}{2\widehat{\sigma}\sqrt{\pi n}},$$

where $\sigma,\widehat{\sigma}$ are the two "variance constants" of Theorem 2.

In summary, as we see in several particular cases here, *qualitatively* similar phenomena are expected in trees, whether plane or non-plane trees, labelled or unlabelled, whereas, *quantitatively*, the structure constants (for instance, μ and $\widehat{\mu}$ in Theorem 2; σ and $\widehat{\sigma}$ in Theorem 4) tend to be model-specific. Yet another instance of such universality phenomena is the height of Otter trees, analysed in [3], which is to be compared to the height of plane binary trees [8]: both scale to $\sqrt{n}$ and lead to the same elliptic-theta distribution, albeit with different scaling factors.

Acknowledgements. The work of M. Bóna was partially supported by the National Science Foundation and the National Security Agency. The work of P. Flajolet was partly supported by the French ANR Project SADA ("Structures Discrètes et Algorithmes").

References

[1] Bender, E. A. Central and local limit theorems applied to asymptotic enumeration. *Journal of Combinatorial Theory 15* (1973), 91–111.

[2] Bóna, M., and Knopmacher, A. On the probability that certain compositions have the same number of parts. *Annals of Combinatorics* (2008). To appear, 19pp.

[3] Broutin, N., and Flajolet, P. The height of random binary unlabelled trees. In *Proceedings of Fifth Colloquium on Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities* (Blaubeuren, 2008), U. Rösler, Ed., vol. AI, pp. 121–134.

[4] Diestel, R. *Graph Theory*. No. 173 in Graduate Texts in Mathematics. Springer Verlag, 2000.
[5] Erdős, P., and Turán, P. On some problems of a statistical group theory III. *Acta Math. Acad. Sci. Hungar. 18* (1967), 309–320.
[6] Finch, S. *Mathematical Constants*. Cambridge University Press, 2003.
[7] Flajolet, P., Fusy, E., Gourdon, X., Panario, D., and Pouyanne, N. A hybrid of Darboux's method and singularity analysis in combinatorial asymptotics. *Electronic Journal of Combinatorics 13*, 1:R103 (2006), 1–35.
[8] Flajolet, P., and Odlyzko, A. M. The average height of binary trees and other simple trees. *Journal of Computer and System Sciences 25* (1982), 171–213.
[9] Flajolet, P., and Sedgewick, R. *Analytic Combinatorics*. Cambridge University Press, 2008. In press; 825 pages (ISBN-13: 9780521898065); also available electronically from the authors' home pages.
[10] Gnedenko, B. V., and Kolmogorov, A. N. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, 1968. Translated from the Russian original (1949).
[11] Goh, W. M. Y., and Schmutz, E. The expected order of a random permutation. *Bulletin of the London Mathematical Society 23*, 1 (1991), 34–42.
[12] Harary, F., and Palmer, E. M. *Graphical Enumeration*. Academic Press, 1973.
[13] Hwang, H.-K. On convergence rates in the central limit theorems for combinatorial structures. *European Journal of Combinatorics 19*, 3 (1998), 329–343.
[14] McKeon, K. A. The expected number of symmetries in locally-restricted trees I. In *Graph Theory, Combinatorics, and Applications*, Y. Alavi, Ed. Wiley, 1991, pp. 849–860.
[15] McKeon, K. A. The expected number of symmetries in locally restricted trees II. *Discrete Applied Mathematics 66*, 3 (1996), 245–253.
[16] Nicolas, J.-L. Distribution statistique de l'ordre d'un élément du groupe symétrique. *Acta Math. Hung. 45*, 1–2 (1985), 69–84.
[17] Otter, R. The number of trees. *Annals of Mathematics 49*, 3 (1948), 583–599.
[18] Pólya, G., and Read, R. C. *Combinatorial Enumeration of Groups, Graphs and Chemical Compounds*. Springer Verlag, 1987.
[19] Sloane, N. J. A. *The On-Line Encyclopedia of Integer Sequences*. 2008. Published electronically at `www.research.att.com/~njas/sequences/`.
[20] Stanley, R. P. *Enumerative Combinatorics*, vol. II. Cambridge University Press, 1999.
[21] Van Cutsem, B., and Ycart, B. Indexed dendrograms on random dissimilarities. *Journal of Classification 15*, 1 (1998), 93–127.
[22] Wilf, H. S. The variance of the Stirling cycle numbers. Tech. rep., ArXiv, 2005.

M. Bóna, Department of Mathematics, University of Florida, 358 Little Hall, PO Box 118105, Gainesville, FL 32611–8105 (USA)

P. Flajolet. Algorithms Project, INRIA Rocquencourt, F-78153 Le Chesnay (France)